Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1-PB2-HA Gene Complex

MBE Advance Access published October 31, 2014

Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1–PB2–HA Gene Complex

Gytis Dudas,*[1] Trevor Bedford,[2] Samantha Lycett,[1,3] and Andrew Rambaut[1,4,5]

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom

[2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA

[3]Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, Glasgow, United Kingdom

[4]Fogarty International Center, National Institutes of Health, Bethesda, MD

[5]Centre for Immunology, Infection and Evolution at the University of Edinburgh, Edinburgh, United Kingdom

***Corresponding author:** E-mail: g.dudas@sms.ed.ac.uk.

Associate editor: Robin Bush

Abstract

Influenza B viruses make a considerable contribution to morbidity attributed to seasonal influenza. Currently circulating influenza B isolates are known to belong to two antigenically distinct lineages referred to as B/Victoria and B/Yamagata. Frequent exchange of genomic segments of these two lineages has been noted in the past, but the observed patterns of reassortment have not been formalized in detail. We investigate interlineage reassortments by comparing phylogenetic trees across genomic segments. Our analyses indicate that of the eight segments of influenza B viruses only segments coding for polymerase basic 1 and 2 (PB1 and PB2) and hemagglutinin (HA) proteins have maintained separate Victoria and Yamagata lineages and that currently circulating strains possess PB1, PB2, and HA segments derived entirely from one or the other lineage; other segments have repeatedly reassorted between lineages thereby reducing genetic diversity. We argue that this difference between segments is due to selection against reassortant viruses with mixed-lineage PB1, PB2, and HA segments. Given sufficient time and continued recruitment to the reassortment-isolated PB1–PB2–HA gene complex, we expect influenza B viruses to eventually undergo sympatric speciation.

Key words: **influenza, reassortment, evolution, phylogenetics, speciation.**

Introduction

Seasonal influenza causes between 250,000 and 500,000 deaths annually and comprises lineages from three virus types (A, B, and C) cocirculating in humans, of which influenza A is considered to cause the majority of seasonal morbidity and mortality (World Health Organization 2009). Occasionally influenza B viruses become the predominant circulating virus in some locations, for example in the 2012/2013 European season as many as 53% of influenza sentinel surveillance samples tested positive for influenza B (Broberg et al. 2013).

Like other members of *Orthomyxoviridae*, influenza B viruses have segmented genomes, which allow viruses coinfecting the same cell to exchange segments, a process known as reassortment. Influenza A viruses are widely considered to be a major threat to human health worldwide due to their ability to cause pandemics in humans through reassortment of circulating human strains with nonhuman influenza A strains. Although influenza B viruses have been observed to infect seals (Osterhaus et al. 2000; Bodewes et al. 2013) through a reverse zoonosis, they are thought to primarily infect humans and are thus unlikely to exhibit pandemics due to the absence of an animal reservoir from which to acquire antigenic novelty. Both influenza A and B evolve antigenically through time in a process known as antigenic drift, in which mutations to the hemagglutinin (HA) protein allow viruses to escape existing human immunity and persist in the human population, leading to recurrent seasonal epidemics (Burnet 1955; Hay et al. 2001; Bedford et al. 2014).

Currently circulating influenza B viruses comprise two distinct lineages—Victoria and Yamagata (referred to as Vic and Yam, respectively)—named after strains B/Victoria/2/87 and B/Yamagata/16/88 that are thought to have genetically diverged in HA around 1983 (Rota et al. 1990). These two lineages now possess antigenically distinct HA surface glycoproteins (Kanegae et al. 1990; Rota et al. 1990; Nerome et al. 1998; Nakagawa et al. 2002; Ansaldi et al. 2003) allowing them to cocirculate in the human population. Phylogenetic analysis of evolutionary rate, selective pressures, and reassortment history of influenza B has shown extensive and often complicated patterns of reassortment between all segments of influenza B viruses both between and within the Vic and Yam lineages (Chen and Holmes 2008).

Here, we extend previous methods to reveal an intriguing pattern of reassortment in influenza B. In our approach, membership to either the Vic or Yam lineage in one segment is used to label the individual isolates in the tree of the other segments. By modeling the transition between labels on a phylogenetic tree, reassortment events which result in the replacement of one segment's lineage by another show up as label changes along a branch (fig. 1). We use this method to reconstruct major reassortment events and quantify reassortment dynamics over time in a data set of 452 influenza B

Article

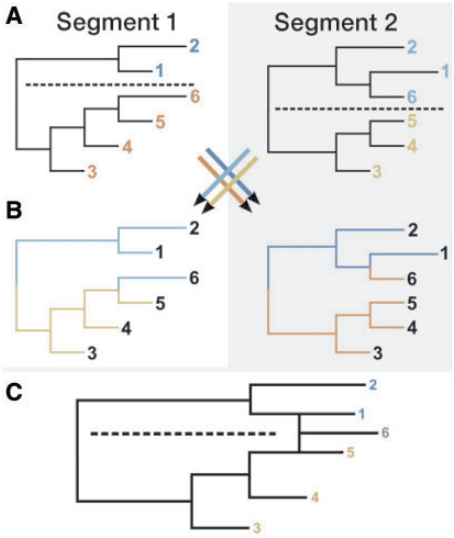

FIG. 1. Schematic analysis of reassortment patterns. (*A*) We begin by assigning sequences falling on either side of a specified bifurcation within each segment tree to different lineages, in this case, the Vic and Yam bifurcation that occurred in the early 1980s. (*B*) We then transfer lineage labels from one tree to the same tips in another tree. Transitions between labels along this second tree thus indicate reassortment events that combine lineages falling on different sides of the Vic/Yam bifurcation in the first tree. (*C*) A reassortment graph depiction shows that tip number 6 is determined to be a reassortant based on (*B*).

genomes, and conduct secondary analyses in a data set of 1,603 influenza B genomes.

We show that despite extensive reassortment, three of the eight segments—two segments coding for components of the influenza B virus polymerase, PB1 and PB2, and the surface glycoprotein HA—still survive as distinct Vic and Yam lineages, which appear to be codependent to the point where virions which do not contain PB1, PB2, or HA segments derived entirely from either the Vic or the Yam lineage have rarely been isolated and only circulate as transient lineages once isolated. In other segments (PA, NP, NA, MP, and NS) a single lineage has introgressed into the opposing background and replaced the previous lineage: All currently circulating influenza B viruses have PA, NP, NA, and MP segments derived from Yam lineage and NS segments derived from Vic lineage.

Results

Analysis of Reassortment Patterns across Vic and Yam Lineages

The differentiation into Vic and Yam lineages can be seen in all segments (fig. 2) and is followed by interlineage reassortment events. In the phylogenetic trees of the PA, NP, NA, MP, and NS segments, either the Vic or Yam lineage has become the "trunk" of the tree, with present-day viruses deriving entirely from the Vic or Yam lineage (yellow vs. purple bars in fig. 2) following reassortment. However, the Vic and Yam lineages of PB1, PB2, and HA segments continue to cocirculate to this day. Periodic loss of diversity in PA, NP, NA, MP, and NS segments is consistent with introgression of one lineage into the other in those segments, whereas maintenance of parallel Vic and Yam lineages results in continually increasing diversity in segments PB1, PB2, and HA (fig. 3). The PB1, PB2, and HA segments from present-day viruses maintain a common ancestor in approximately 1983 and thus accumulate genetic diversity since the split of those segments into Vic and Yam lineages, whereas other segments often lose diversity with ancestors to present-day viruses appearing between approximately 1991 and approximately 1999.

By measuring mean pairwise diversity between branches in each tree that were assigned either a Vic or Yam label in other segments, we look for reductions in between-lineage diversity, which indicate that an interlineage reassortment event has taken place (fig. 4). This method gives a quantitative measure of reassortment-induced loss of diversity between Vic and Yam lineages in two trees, although care should be taken when interpreting the statistic, as it does not correspond to any real time of most recent common ancestors (TMRCAs) in the tree, but can be interpreted as mean coalescence date between Vic and Yam lineages of PB1, PB2, and HA segments in all other trees. We focus only on PB1, PB2, and HA lineage labels, as all other segments eventually become completely derived from either the Vic or the Yam lineage. Losses of diversity (represented by more recent mean pairwise TMRCAs between Vic and Yam labels) in figure 4 indicate that every segment has reassorted with respect to the Vic and Yam lineages of PB1, PB2, and HA segments. However, we also see that the labels for these three segments show reciprocal preservation of diversity after 1997. This suggests that after 1997 no reassortment events have taken place between Vic and Yam lineages of PB1, PB2, and HA segments and their lineage labels only "meet" at the root. We do see reduced diversity between Vic and Yam labels of PB1, PB2, and HA segments in a time period close to the initial split of Vic and Yam lineages (1986–1996). These reductions in diversity represent small clades with reassortant PB1–PB2–HA constellations, which go extinct by 1997 (see fig. 2). We also observe that the assignment of these three segment labels to branches of other segment trees is very similar and often identical after 1997. This suggests that PB1, PB2, and HA lineage labels switch simultaneously in all trees after 1997.

We show the ratio of Vic to Yam sequences in our primary and secondary data sets in different influenza seasons in figure 5, which is based on which lineage each sequence was assigned to (see Materials and Methods). It is evident that losses of diversity in the PA, NP, NA, MP, and NS segments are related to either the Vic (NS) or the Yam (PA, NP, NA, and MP) lineage replacing the other lineage in the influenza B virus population. Similarly, the lack of reassortment between Vic and Yam lineages and maintenance of diversity of PB1, PB2, and HA can be seen, where the two lineages have been sequenced at a ratio close to 50% over long periods of time (fig. 5). On a year-to-year basis, however, the ratios for Vic and Yam sequences PB1, PB2, and HA can fluctuate dramatically consistent with one lineage predominating within a given season, in agreement with surveillance data (Reed et al. 2012).

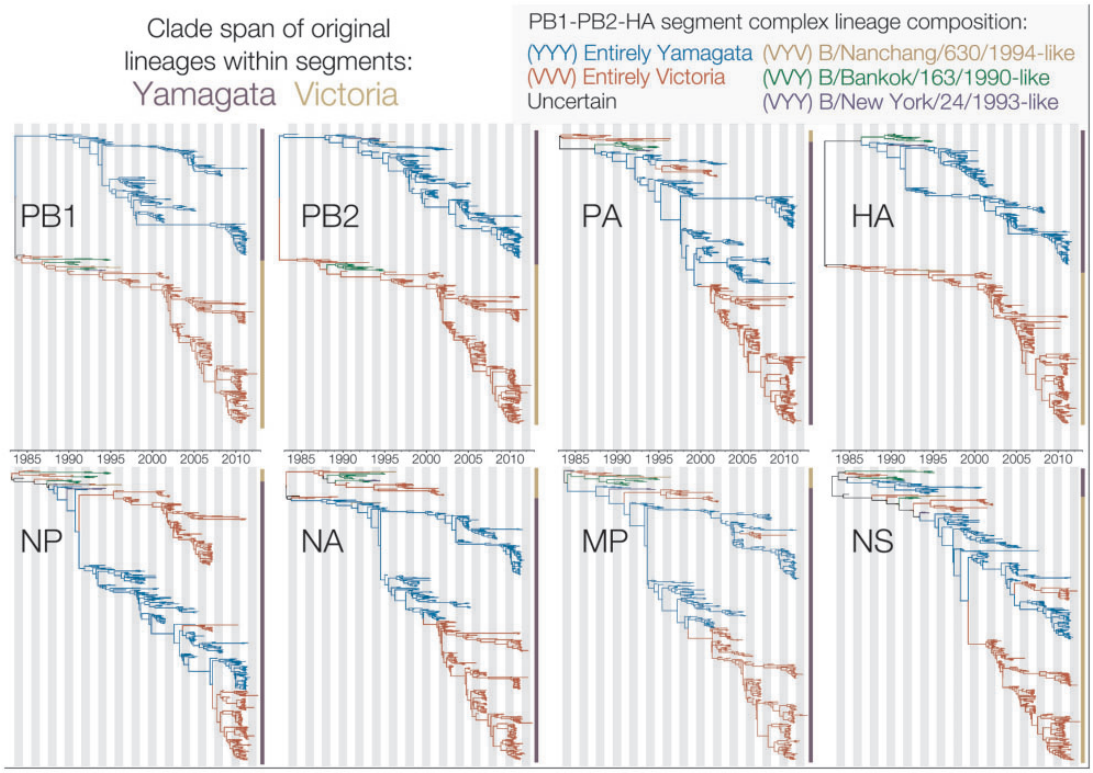

Fig. 2. Maximum clade credibility trees of all eight genome segments of influenza B viruses isolated since 1980. Trees are colored based on inferred PB1–PB2–HA lineage. Vertical bars indicate the original Vic and Yam lineages within each segment. Each tree is the summarized output of a single analysis comprised of 9,000 trees sampled from the posterior distribution of trees.

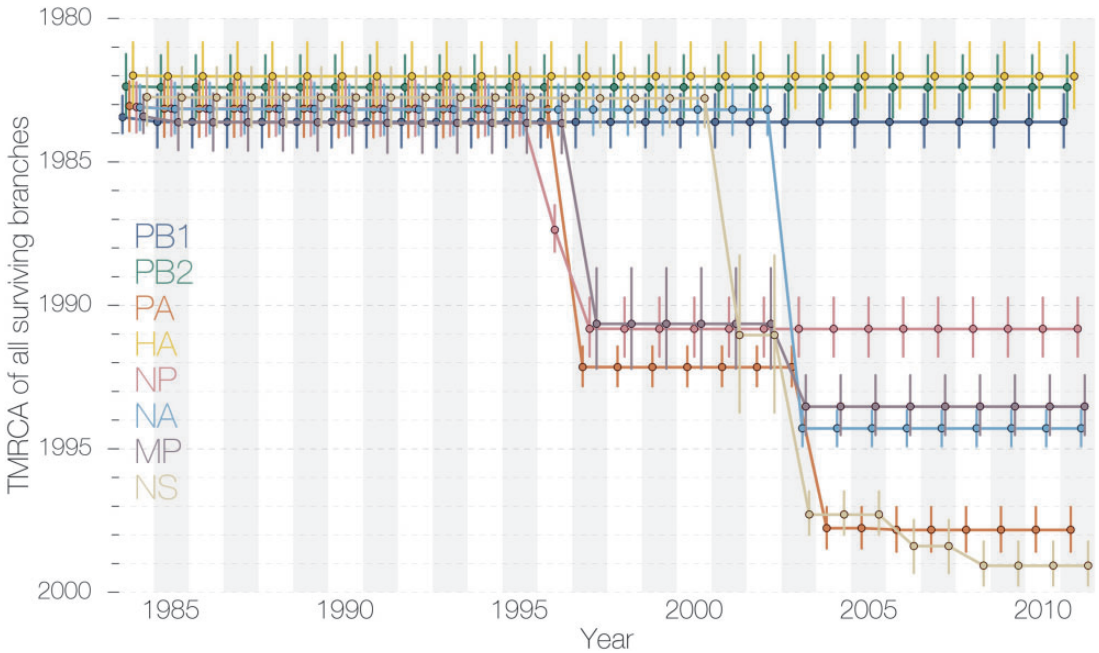

Fig. 3. Oldest TMRCA of all surviving branches over time. PA, NP, NA, MP, and NS segments of influenza B viruses show periodic increases in TMRCAs of all surviving branches (indicative of diversity loss), suggesting lineage turnover. PB1, PB2, and HA segments, on the other hand, maintain the diversity dating back to the initial split of Vic and Yam lineages. Each point is the mean TMRCA of all surviving lineages existing at each time slice through the tree and vertical lines indicating uncertainty are 95% highest posterior densities.

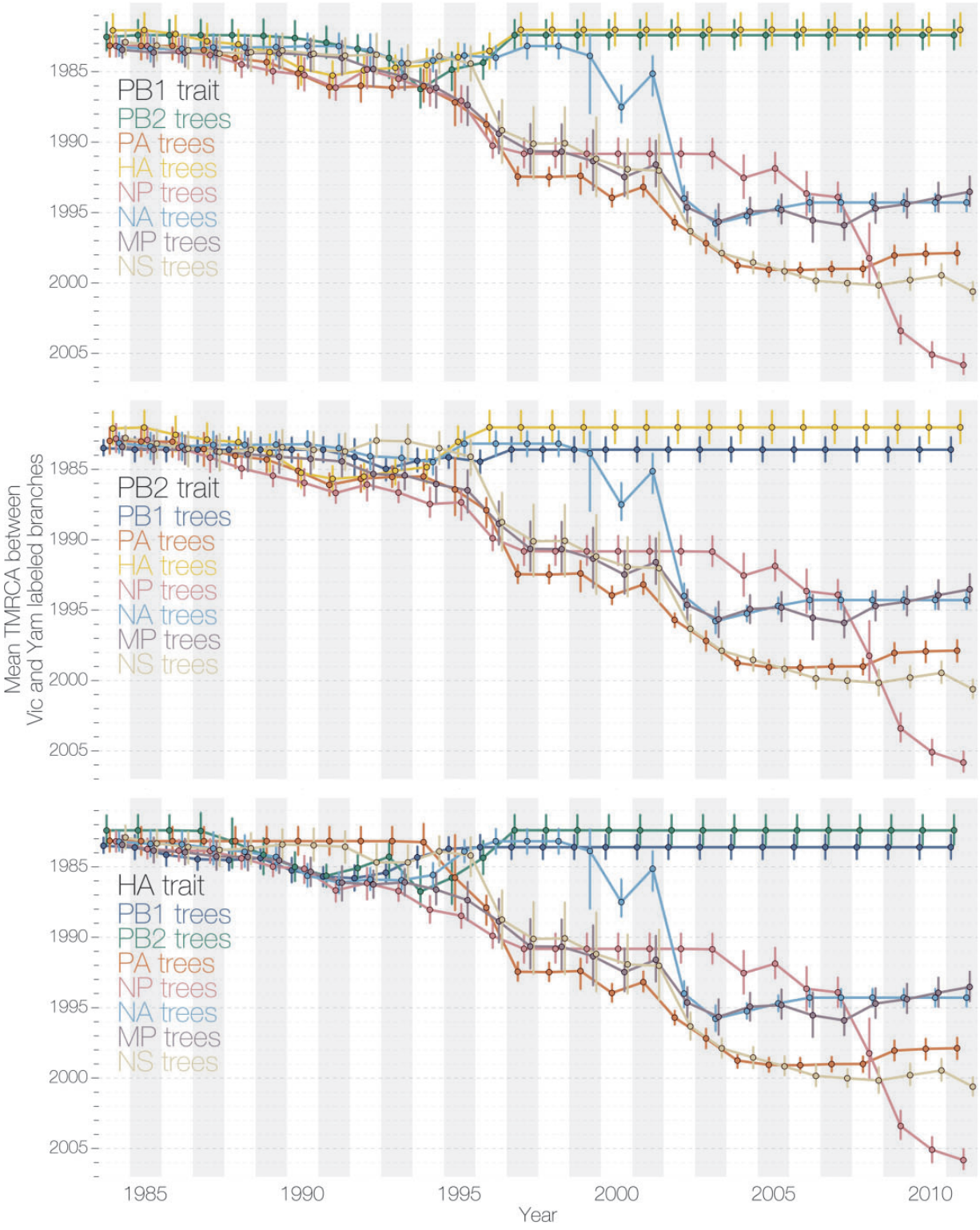

FIG. 4. Mean pairwise TMRCA between Vic and Yam branches under PB1, PB2, and HA label sets. PB1, PB2, and HA segment labels indicate that these segments show reciprocal preservation of diversity, which dates back to the split of Vic and Yam lineages. All other segments show increasingly more recent TMRCAs between branches labeled as Vic and Yam in PB1, PB2, and HA label sets. All vertical lines indicating uncertainty are 95% highest posterior densities.

We reconstructed reassortment events that were detected by using lineage labels. Figure 6 focuses only on interlineage reassortments that have occurred after 1990. We identify five major (in terms of persistence) reassortant genome constellations (given in order PB1–PB2–PA–HA–NP–NA–MP–NS with prime [′] indicating independently acquired segments) circulating between 1992 and 2011 (fig. 6):

- B/Alaska/12/1996-like (Y–Y–Y–Y–Y–Y–Y–V)
- B/Nanchang/2/1997-like (V–V–Y–V–Y–V–Y–V)

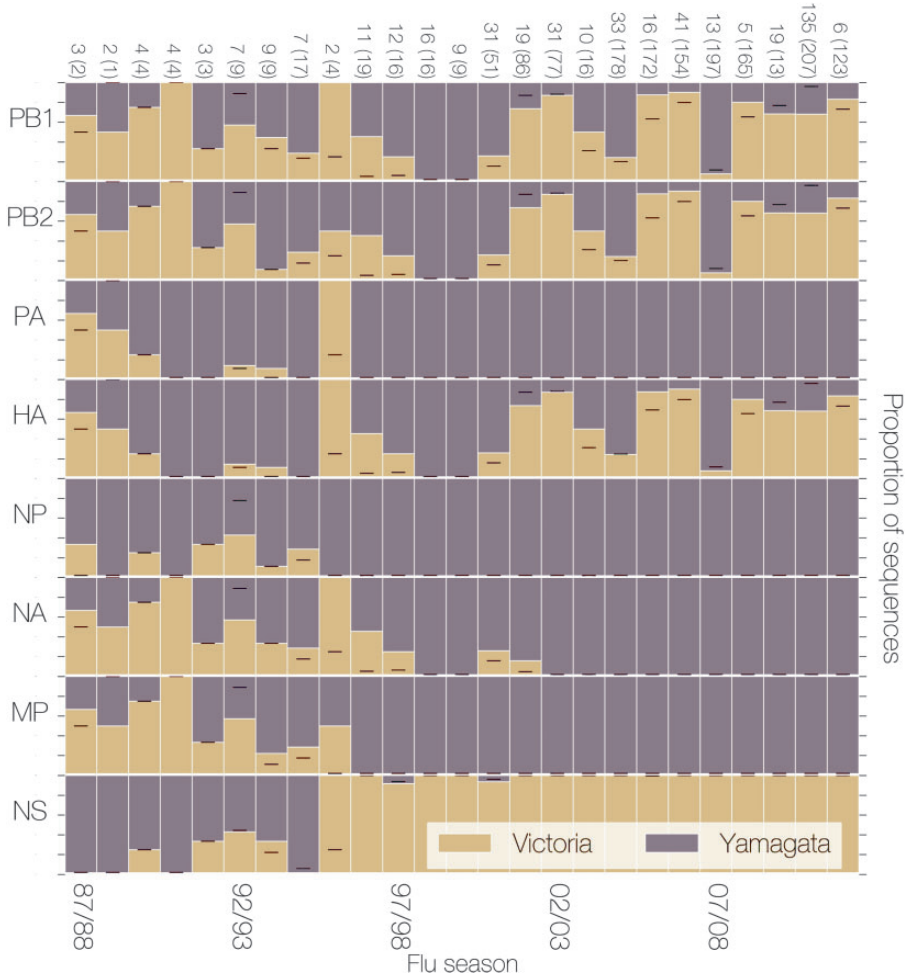

Fig. 5. Ratio of Vic and Yam sequences in the data set. The ratio of Vic (yellow) to Yam (purple) sequences in each segment from the primary data set over time. Black lines indicate where this ratio lies in the larger secondary data set. Numbers at the top of the figure show the total number of genomes available for each influenza season in the primary data set comprised of 452 genomes from which the ratio was calculated, whereas the numbers in brackets correspond to numbers of sequences in the larger secondary genomes data set. Numbers at the bottom are influenza seasons from the 1987/1988 (87/88) season to the 2011/2012 season. Yam lineage of PA, NP, NA, and MP segments and Vic lineage of the NS segment eventually become fixed (in the population genetics sense of the word) in the influenza B population. PB1, PB2, and HA segments maintain separate Vic and Yam lineages.

- B/Iowa/03/2002-like (V–V–Y′–V–Y–Y–Y′–V′)
- B/California/NHRC0001/2006-like (V–V–Y–V–Y′–Y–Y′–V′)
- B/Brisbane/33/2008-like (V–V–Y–V–Y′–Y–Y–V)

In a previous study, B/Alaska/12/1996-like, B/Nanchang/2/1997-like, and B/Iowa/03/2002-like constellations were observed (Chen and Holmes 2008), but sequences from B/California/NHRC0001/2006-like and B/Brisbane/33/2008-like constellations were not available at the time. In their study, Chen and Holmes (2008) also recovered the coassortment pattern of PB1, PB2, and HA lineages, but did not remark upon it. Of these five constellations, four (B/Nanchang/2/1997-like, B/Iowa/03/2002-like, B/California/NHRC0001/2006-like, and B/Brisbane/33/2008-like) are derived from introgression of Yam lineage segments into Vic lineage PB1–PB2–HA background, with only one (B/Alaska/12/1996-like) resulting from introgression of Vic lineage NS segment into an entirely Yam-derived background. All five interlineage reassortment events described here are marked by the preservation of either entirely Vic- or Yam-derived PB1–PB2–HA segments. Figure 6 also shows that reassorting segments appear to evolve with a considerable degree of autonomy. For example, the NP lineage that entered a largely Vic lineage-derived genome and gave rise to the B/Nanchang/2/1997-like isolates continued circulating until 2010, even though the other segments it coassorted with in 1995–1996 (PA and MP) went extinct following the next round of reassortment that led to the rise of B/Iowa/03/2002-like genome constellations. A more extreme example is the NS segment, where a Vic sublineage was reassorted into an entirely Yam background (B/Alaska/12/1996-like) in 1994–1995, then reassorted back into a mostly Vic background some 5 years later (B/Iowa/03/2002-like) where it has replaced the "original" Vic sublineage (see fig. 6).

We observe that in all five successful interlineage reassortment events shown in figure 6, none break up the PB1–PB2–HA complex. This is an unlikely outcome—the probability of not breaking up PB1–PB2–HA across five reassortment events

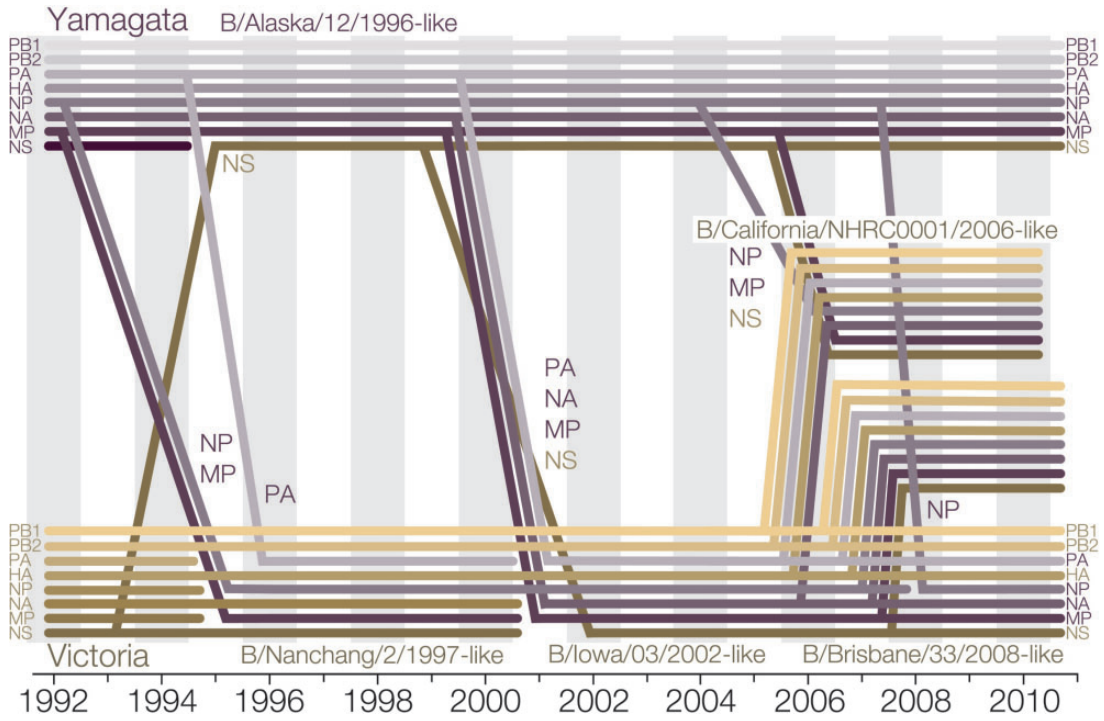

FIG. 6. Schematic plot of reconstructed reassortments between Vic and Yam lineage segments of influenza B virus. Lineages that coassort in genomes are represented by eight parallel lines, with lineages that derive from the original Vic clade colored yellow/brown and lineages that derive from the original Yam clade colored lilac/purple. Interlineage reassortment events are indicated by lines entering a different genome. The angle of incoming lineages represents uncertainty in the timing of the event (mean date of the reassortant node and its parent node). Lineage extinction dates are not shown accurately.

is $P = (\frac{2^5 \times 2 - 2}{2^8 - 2})^5 = 0.0009$, where reassortment events are considered to sample from the Vic and Yam lineages at random for each of the eight segments. If we correct for multiple testing with the assumption that coassortment of any three segments is of interest, we find that the probability of not breaking up an arbitrary set of three segments across five reassortment events is $P = \binom{8}{3} \times (\frac{2^5 \times 2 - 2}{2^8 - 2})^5 = 0.0485$.

Although the vast majority of influenza B isolates possess either Vic or Yam lineage-derived PB1–PB2–HA complexes, on rare occasions mixed-lineage PB1–PB2–HA constellations emerge. Figure 7 shows the sum of branch lengths which were labeled as having entirely Vic, entirely Yam, or mixed-lineage PB1, PB2, and HA segments. Due to lack of reassortment between Vic and Yam lineages of PB1, PB2, and HA (fig. 4), since 1997 all segments have spent significantly longer periods of evolutionary time with either entirely Vic-derived or entirely Yam-derived than with mixed-lineage PB1, PB2, and HA constellations (fig. 7). We have identified three instances of mixed-lineage PB1–PB2–HA reassortants from the primary data set with the following PB1–PB2–HA constellations: VVY (B/Bangkok/163/1990-like, 13 sequences isolated 1990–January 5, 1995), VYV (B/Nanchang/630/1994-like, two sequences isolated 1994–1996), and VYY (B/New York/24/1993-like, two sequences isolated January 8, 1993–1994). We detected two new reassortant lineages when investigating the larger secondary data set—B/Waikato/6/2005-like viruses with PB1–PB2–HA constellation YYV (17 sequences isolated

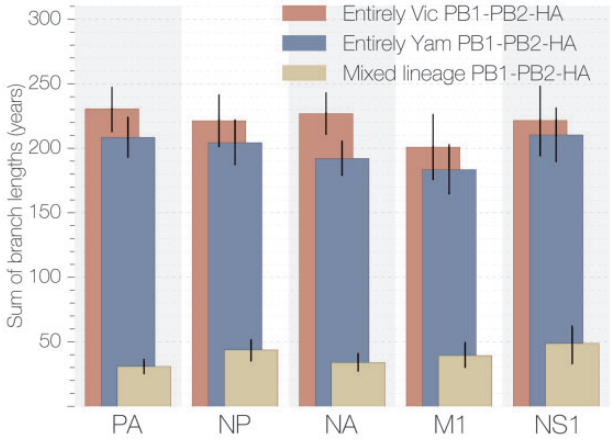

FIG. 7. Amount of evolutionary time each segment has spent under different PB1–PB2–HA constellations. All segments have spent significantly more of their history with entirely Vic- or entirely Yam-derived PB1–PB2–HA complexes. All vertical lines indicating uncertainty are 95% highest posterior densities.

May 9–October 12 in 2005) and B/Malaysia/1829782/2007 with PB1–PB2–HA constellation YVY (one sequence isolated August 2, 2007).

Analysis of Reassortment Properties

We attempted to quantify the temporal discordance between lineages reassorting into new genomic constellations. If one

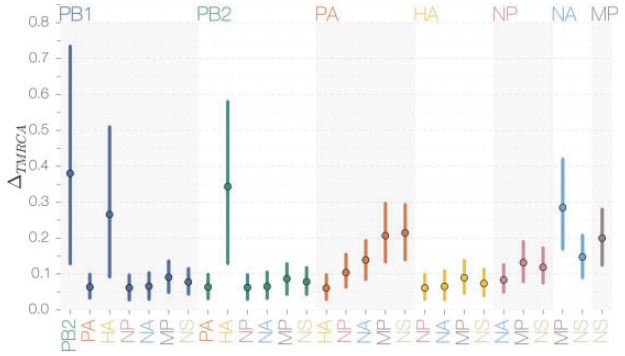

FIG. 8. Δ_{TMRCA} statistics for different segment pairs. PB1, PB2, and HA trees exhibit reciprocally highly similar TMRCAs, unlike most other pairwise comparisons. All vertical lines indicating uncertainty are 95% highest posterior densities.

were able to recover an influenza "species tree," including admixture/reassortment events, it would be possible to estimate the reassortment or recombination "distance," which is the time between a split in the species tree in the past and a reassortment event (see supplementary fig. S17, Supplementary Material online). Although we do not find evidence of differences in total number of reassortments between segments (see supplementary fig. S4, Supplementary Material online), we find support for a reassortment "distance" effect, in which a pair of tips on one segment has a different TMRCA from the same pair of tips on a different segment. The summary statistic we use that reflects this difference in TMRCAs, δ_{TMRCA}, is most sensitive when only one of the two trees being compared loses diversity through reassortment and the other acts like a proxy for the "species tree." We normalize our δ_{TMRCA} comparisons to arrive at Δ_{TMRCA}, which accounts for uncertainty in tree topology (see Materials and Methods). Figure 8 shows Δ_{TMRCA} values for all pairs of trees. Most segment pairs show very low values for this statistic with $\Delta_{TMRCA} \approx 0.1$, indicating that δ_{TMRCA} measurements between replicate posterior samples from the same segment are up to ten times smaller than δ_{TMRCA} values between posterior samples from different segments. PB1, PB2, and HA trees, on the other hand, exhibit Δ_{TMRCA} values that are much higher. This shows that TMRCA differences between trees of PB1, PB2, and HA segments are, though noisy, occasionally very similar to uncertainty in tip-to-tip TMRCAs between replicate analyses of these segments.

Discussion

Linkage between PB1, PB2, and HA Gene Segments

In this article, we show that the PB1, PB2, and HA segments of influenza B viruses are the only ones that have continuously maintained separate Vic and Yam lineages, whereas other segments have fixed either Vic or Yam lineages (figs. 2, 5, and 6). Evidence suggests that this is a result of prolonged lack of reassortment between Vic and Yam lineages in PB1, PB2, and HA (fig. 4), which possess coassorting sequences detectable as high linkage disequilibrium (LD) (supplementary fig. S1, Supplementary Material online). The vast majority of the sampled evolutionary history of each segment of influenza B viruses since the split of Vic and Yam lineages has been spent in association with either completely Vic or completely Yam lineage-derived PB1–PB2–HA complexes (fig. 7), suggesting that having "pure" lineage PB1–PB2–HA complexes is important for whole-genome fitness. We propose that this pattern of coassortment is due to the action of selection and not simply biased or rare reassortment.

The origin of the strong genetic linkage between PB1, PB2, and HA segments remains unclear. We believe that there are two alternative, but similar explanations for the origins of the strong genetic linkage between these segments: Mutation-driven coevolution (Presgraves 2010) and Dobzhansky–Muller (DM) incompatibility (Dobzhansky 1937; Muller 1942). Mutation-driven coevolution (Presgraves 2010) has been suggested to be the cause of hybrid dysfunction in *Saccharomyces* hybrids (Lee et al. 2008) and evolves as a byproduct of adaptation. If one or the other influenza B lineage has undergone adaptation, we might expect these changes to be beneficial in its native background and incompatible with a foreign background. DM incompatibility operates in a similar way, but the main difference from the scenario described earlier is that the incompatible alleles are neutral or nearly neutral in their native background and become deleterious or lethal when combined with nonnative backgrounds. Emergence of DM incompatibility is aided by geographic isolation. Interestingly, the Vic lineage of HA was restricted to eastern Asia between 1992 and 2000 (Nerome et al. 1998; Shaw et al. 2002), offering ample time for the budding Vic lineage to accumulate alleles causing reassortment incompatibility. However, without more genomic data from the past, it is difficult to estimate to what extent influenza B virus population structure contributed to the development of the current segment linkage.

Potential Mechanisms for Reassortment Incompatibility

Unfortunately, the limited amount of genomic data available for the early years of the Vic–Yam split precludes any attempts of answering whether selection or drift have led to the current linkage of PB1, PB2, and HA segments. Although the origins of the linkage between these three segments might be difficult to explain, we can speculate on the nature of reassortment incompatibility. For example, it is intuitive for why this might be the case for PB1 and PB2: Both proteins interact directly as part of the RNA-dependent RNA polymerase heterotrimer. Indeed, we observe that PB1–PB2 reassortants are the rarest and least persistent among mixed-lineage PB1–PB2–HA strains and have not been isolated in great numbers. In fact, most reassortants breaking the PB1–PB2–HA complex apart have occurred in the past, close to the split of Vic and Yam lineages, and have become very rare since.

There is some evidence that the linkage between PB1 and HA might not be a phenomenon restricted to influenza B viruses. It has been established that at least for the 1957 and the 1968 influenza pandemics, caused by A/H2N2 and A/H3N2 subtypes, respectively, the viruses responsible were

reassortants possessing PB1 and HA segments derived from avian influenza A viruses (Kawaoka et al. 1989). In addition, outdated techniques for producing vaccine seed strains through selection for HA–NA reassortants often yielded PB1–HA–NA reassortants as a side-effect (Bergeron et al. 2010; Fulvini et al. 2011). Recent experiments have found that the presence or absence of a "foreign" PB1 segment can have dramatic effects on HA concentration on the surface of virions and total virion production (Cobbin et al. 2013). However, there have been reassortant influenza A viruses circulating for prolonged periods of time in humans that did have disparate PB1 and HA segments, for example, H1N2 outbreaks in 2001 (Gregory et al. 2002) and H1N1/09 in 2009 (Smith et al. 2009).

We believe that the association between PB1, PB2, and HA segments should be relatively straightforward to explore in the lab. Reverse genetics systems have been developed for influenza B viruses (Hoffmann et al. 2002), which would allow the creation of artificial reassortants. Based on the frequency and persistence times of different reassortant classes we have observed, we expect a hierarchy of reassortant fitness starting with PB1+PB2+HA reassortants which should be the most fit, followed by PB1+2/HA, then PB1+HA/PB2, and finally PB2+HA/PB1 reassortants with the lowest fitness. We believe that this is the most direct approach to unraveling the mechanism responsible for the linkage within the PB1–PB2–HA complex.

Will Influenza B Viruses Speciate?

We suggest that the preservation of two PB1–PB2–HA complex lineages is similar to genomic speciation islands, where small numbers of genes resist being homogenized through gene flow (Turner et al. 2005). In this context, we see three potential paths of evolution for influenza B viruses. If more segments get recruited to the PB1–PB2–HA complex, the process could continue until "speciation" occurs in which none of the segments is able to reassort across the Vic–Yam lineage boundary. Alternatively, the influenza B genome could continue to be homogenized through gene flow with the exception of PB1, PB2, and HA segments or one of the two PB1–PB2–HA complexes could go extinct, marking the return of single-strain dynamics in the influenza B virus population. The eventual fate of influenza B viruses will likely be determined by the combined effects of reassortment frequency and the strength of epistatic interactions between segments.

Materials and Methods

We compiled a primary data set of 452 complete influenza B genomes from GISAID (Bogner et al. 2006) dating from 1984 to 2012. The longest protein-coding region of each segment was extracted and used for all further analyses. We thus assume that homologous recombination has not taken place and that the evolutionary history of the whole segment can be inferred from the longest coding sequence in the segment. To date, there has been little evidence of homologous recombination in influenza viruses (Chare et al. 2003; Boni et al. 2008; Han et al. 2010). The segments of each strain were assigned to either Vic or Yam lineage by making maximum-likelihood trees of each segment using PhyML (Guindon and Gascuel 2003) and identifying whether the isolate was more closely related to B/Victoria/2/87 or B/Yamagata/16/88 sequences in that segment, with the exception of the NS segment as B/Victoria/2/87 was a reassortant and possessed a Yam lineage NS (Lindstrom et al. 1999). B/Czechoslovakia/69/1990 was considered as being representative of Vic lineage for the NS segment. Every segment in each genome thus received either a Vic or a Yam lineage designation, for example, the strain B/Victoria/2/87 received V–V–V–V–V–V–V–Y, as its NS segment is derived from the Yam lineage and the rest of the genome is Vic.

We also collated a secondary data set from all complete influenza B virus genomes available on GenBank as of May 5, 2014. After removing isolates that had considerable portions of any sequence missing, were isolated prior to 1980, or were suspected of having a contaminant sequence in any segment, we were left with 1,603 sequences. This data set only became available after all primary analyses were performed, are mainly from Australia, New Zealand and the United States, and are too numerous to analyze in BEAST (Drummond et al. 2012). PhyML (Guindon and Gascuel 2003) was used to produce phylogenies of each segment, and the lineage of each isolate was determined based on grouping with either B/Victoria/2/87 or B/Yamagata/16/88 sequences, as described above. By associating strains with lineage identity of each of their segments, we reconstructed the most parsimonious interlineage reassortment history for the secondary data set. The secondary data set was used to check how representative the primary data set was, to estimate LD, and to broadly confirm our results. All analyses pertain to the primary data set unless stated otherwise.

Temporally calibrated phylogenies were recovered for each segment in the primary data set using Markov chain Monte Carlo (MCMC) methods in the BEAST software package (Drummond et al. 2012). We modeled the substitution process using the Hasegawa–Kishino–Yano model of nucleotide substitution (Hasegawa et al. 1985), with separate transition models for each of the three codon partitions, and additionally estimated realized synonymous and nonsynonymous substitution counts (O'Brien et al. 2009). We used a flexible Bayesian skyride demographic model (Minin et al. 2008). We accounted for incomplete sampling dates for 94 sequences (of which 93 had only year and 1 had only year and month of isolation) whereby tip date is estimated as a latent variable in the MCMC integration. A relaxed molecular clock was used, where branch rates are drawn from a lognormal distribution (Drummond et al. 2006). We ran three independent MCMC chains, each with 200 million states, sampled every 20,000 steps and discarded the first 10% of the MCMC states as burn-in. After assessing convergence of all three MCMC chains by visual inspection using Tracer (Rambaut et al. 2009), we combined samples across chains to give a total of 27,000 samples from the posterior distribution of trees.

Every sequence was assigned seven discrete traits in BEAUti corresponding to the lineages of all other segments with

which a strain was isolated, for example, PB1 tree had PB2, PA, HA, NP, NA, MP, and NS as traits and V or Y as values for each trait. We inferred the ancestral state of lineages in each segment by modeling transitions between these discrete states using an asymmetric transition matrix (Lemey et al. 2009) with Bayesian stochastic search variable selection to estimate significant rates. Because the posterior set of trees for a single segment has branches labeled with the inferred lineage in the remaining seven segments, we can detect interlineage reassortments between pairs of segments by observing state transitions, that is, Yam to Vic or Vic to Yam (fig. 1). In addition, by reconstructing the ancestral state of all other genomic segments jointly we can infer coreassortment events when more than one trait transition occurs on the same node in a tree. Interphylogeny labeling approaches have been extensively used in the past to investigate reticulate evolution in influenza A viruses and HIV (Lycett et al. 2012; Ward et al. 2013; Lu et al. 2014).

Measures of Diversity

We inferred the diversity of each segment from their phylogenetic tree by estimating the date of the most recent common ancestor of all branches at yearly time points, which places an upper bound on the maximum amount of diversity existing at each time point. A version of this lineage turnover metric has previously been used to investigate the tempo and strength of selection in influenza A viruses during seasonal circulation (Bedford et al. 2011). In addition, we calculated mean pairwise TMRCA between branches labeled as Vic and Yam for PB1, PB2, and HA traits. This gave us a measure of how much a particular segment reassorts with respect to Vic and Yam lineages of PB1, PB2, and HA segments. If Vic and Yam lineages of PB1, PB2, and HA segments were to be considered as being separate populations, this measure would be equivalent to "between population" diversity.

We also calculated the total amount of sampled evolutionary time spent by each segment with entirely Vic, entirely Yam, or mixed-lineage PB1, PB2, and HA segments. We do this by summing the branch lengths in each tree under three different lineage combinations of the PB1, PB2, and HA segments: PB1–PB2–HA derived entirely from Yam lineage, PB1–PB2–HA entirely derived from Vic lineage, and PB1–PB2–HA derived from a mixture of the two lineages. This gives a measure of how successful, over long periods of time, each particular PB1–PB2–HA constellation has been.

Tree-to-Tree Similarities

We express the normalized distance Δ_{TMRCA} between trees belonging to two segments A and B for a particular posterior sample i, following

$$\Delta_{\mathrm{TMRCA}}(A_i, B_i) = \frac{\delta_{\mathrm{TMRCA}}(A_i, A_i') + \delta_{\mathrm{TMRCA}}(B_i, B_i')}{2\,\delta_{\mathrm{TMRCA}}(A_i, B_i)}, \quad (1)$$

where $\delta_{\mathrm{TMRCA}}(A_i, B_i) = \frac{1}{n}\sum_{j=1}^{n} g(A_{ij}, B_{ij})$ and n is the total number of pairwise comparisons available between sets of tips. Thus, $g(A_{ij}, B_{ij})$ is the absolute difference in TMRCA of a pair of tips j, where the pair is drawn from the ith posterior sample of tree A and the ith posterior sample of tree B. Additionally, $\delta_{\mathrm{TMRCA}}(A_i, A_i')$ is calculated from the ith posterior sample of tree A and ith posterior sample of an independent analysis of tree A (which we refer to as A'), which is used in the normalization procedure to control for variability in tree topology stability over the course of the MCMC chain (see supplementary figs. S6 and S7, Supplementary Material online). We had three replicate analyses of each segment and in order to calculate $\delta_{\mathrm{TMRCA}}(A_i, A_i')$ we used analyses numbered 1, 2 and 3 as A and analyses numbered 2, 3 and 1 as A', in that order. We subsampled our combined posterior distribution of trees to give a total of 2,700 trees on which to analyze Δ_{TMRCA}.

Calculating the normalized $\Delta_{\mathrm{TMRCA}}(A_i, B_i)$ for each MCMC state provides us with a posterior distribution of this statistic allowing specific hypotheses regarding similarities between the trees of different segments to be tested. Our approach exploits the branch scaling used by BEAST (Drummond et al. 2012), as the trees are scaled in absolute time and insensitive to variation in nucleotide substitution rates between segments, allowing for direct comparisons between TMRCAs in different trees. In the absence of reassortment we expect the tree of every segment to recapitulate the "virus tree," a concept analogous to "species trees" in population genetics. Our method operates under the assumption that the segment trees capture this "virus tree" of influenza B viruses quite well. It is not an unreasonable assumption, given the seasonal bottlenecks influenza viruses experience. This makes it almost certain that influenza viruses circulating at any given time point are derived from a single genome that existed in the recent past. The δ_{TMRCA} statistic essentially quantifies the temporal distance between admixture events and nodes in the virus tree (see supplementary fig. S17, Supplementary Material online). We normalize δ_{TMRCA} values to get Δ_{TMRCA}, a measure which quantifies the extent to which the similarity of two independent trees resembles phylogenetic noise. The δ_{TMRCA} statistic is an extension of patristic distance methods and has previously been used to tackle a wide variety of problems, as phylogenetic distance in predicting viral titer in *Drosophila* infected with viruses from closely related species (Longdon et al. 2011), and to assess temporal incongruence in a phylogenetic tree of amphibian species induced by using different calibrations (Ruane et al. 2011).

LD across the Influenza B Genome

We used the secondary GenBank data set with 1,603 complete genome sequences to estimate LD between amino acid loci across the longest proteins encoded by each segment of the influenza B virus genome. To quantify LD, we adapt the χ^2_{df} statistic from (Hedrick and Thomson 1986):

$$\chi^2_{\mathrm{df}} = \frac{\chi^2}{N\,(k-1)\,(m-1)}, \quad (2)$$

where χ^2 is calculated from a classical contingency table, N is the number of haplotypes, and $(k-1)(m-1)$ are the

degrees of freedom. This statistic is equal to the widely used r^2 LD statistic at biallelic loci, but also quantifies LD when there are more than two alleles per locus (Zhao et al. 2005). LD was estimated only at loci where each nucleotide or amino acid allele was present in at least two isolates. We ignored gaps in the alignment and did not consider them as polymorphisms. In all cases, we used a minor allele frequency cutoff of 1%. We also calculated another LD statistic, D' (Lewontin 1964) as $D'_{ij} = D_{ij}/D^{\max}_{ij}$, where $D_{ij} = p(A_iB_j) - p(A_i)p(B_j)$ and

$$\begin{aligned} D^{\max}_{ij} &= \min[p(A_i)p(B_j), (1-p(A_i))(1-p(B_j))] \\ &\text{when } D_{ij} < 0 \\ D^{\max}_{ij} &= \min[(1-p(A_i))p(B_j), p(A_i)(1-p(B_j))] \\ &\text{when } D_{ij} \geq 0, \end{aligned} \tag{3}$$

where $p(A_i)$ is the frequency of allele A_i at locus A, $p(B_j)$ is the frequency of allele B_j at locus B, and $p(A_iB_j)$ is the frequency of haplotype A_iB_j. D' is inflated when some haplotypes are not observed, for example, when the minor allele frequency is low. We find that D' is almost uniformly high across the influenza B virus genome and close to 1.0 for almost any pair of polymorphic loci. This is because most amino acid alleles in the population exist transiently, meaning that they do not get a chance to reassort and we only observe them within the backgrounds of more persistent alleles, which D' quantifies as complete LD. We think that metrics related to r^2, such as χ^2_{df}, perform much better on temporal data such as ours in finding persistent associations between alleles and are easier to interpret.

Data Availability

Python scripts used to process trees and sequences are available at https://github.com/evogytis/fluB/tree/master/scripts (last accessed October 21, 2014). Output files from scripts, lineage designations, MCC trees, acknowledgment tables, accession numbers, and redacted XML files (per GISAID Data Access Agreement) are publicly available at https://github.com/evogytis/fluB/tree/master/data (last accessed October 21, 2014).

Supplementary Material

Supplementary material and figures S1–S17 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Darren Obbard and Paul Wikramaratna for helpful discussions and anonymous reviewers for comments and suggestions. This study was supported by a Natural Environment Research Council studentship D76739X to G.D. and a Newton International Fellowship from the Royal Society to T.B. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 278433-PREDEMICS and ERC Grant Agreement No. 260864. A.R. and S.L. acknowledge the support of the Wellcome Trust (Grant No. 092807).

References

Ansaldi F, D'Agaro P, de Florentiis D, Puzelli S, Lin YP, Gregory V, Bennett M, Donatelli I, Gasparini R, Crovari P, et al. 2003. Molecular characterization of influenza B viruses circulating in northern Italy during the 2001–2002 epidemic season. *J Med Virol* 70:463–469.

Bedford T, Cobey S, Pascual M. 2011. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evol Biol.* 11:220.

Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A.. 2014. Integrating influenza antigenic dynamics with molecular evolution. *eLife* 3: p. e01914.

Bergeron C, Valette M, Lina B, Ottmann M. 2010. Genetic content of influenza H3N2 vaccine seeds. *PLoS Curr.* 2:RRN1165.

Bodewes R, Morick D, de Mutsert G, Osinga N, Bestebroer T, van der Vliet S, Smits SL, Kuiken T, Rimmelzwaan GF, Fouchier RA, et al. 2013. Recurring influenza B virus infections in seals. *Emerg Infect Dis.* 19:511–512.

Bogner P, Capua I, Lipman DJ, Cox NJ, et al. 2006. A global initiative on sharing avian flu data. *Nature* 442:981.

Boni MF, Zhou Y, Taubenberger JK, Holmes EC. 2008. Homologous recombination is very rare or absent in human influenza A virus. *J Virol.* 82:4807–4811.

Broberg E, Beauté J, Snacken R 2013. Fortnightly influenza surveillance overview, 24 May 2013 - weeks 19-20/2013. Available from: http://ecdc.europa.eu/en/publications/Publications/influenza-fortnightly-surveillance-overview-24-may-2013.pdf.

Burnet SFM. 1955. Principles of animal virology. New York: Academic Press.

Chare ER, Gould EA, Holmes EC. 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *J Gen Virol.* 84:2691–2703.

Chen R, Holmes EC. 2008. The evolutionary dynamics of human influenza B virus. *J Mol Evol.* 66:655–663.

Cobbin JCA, Verity EE, Gilbertson BP, Rockman SP, Brown LE. 2013. The source of the PB1 gene in influenza vaccine reassortants selectively alters the hemagglutinin content of the resulting seed virus. *J Virol.* 87:5577–5585.

Dobzhansky T. 1937. Genetics and the origin of species. New York: Columbia University Press.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29: 1969–1973.

Fulvini AA, Ramanunninair M, Le J, Pokorny BA, Arroyo JM, Silverman J, Devis R, Bucher D. 2011. Gene constellation of influenza A virus reassortants with high growth phenotype prepared as seed candidates for vaccine production. *PLoS One* 6:e20823.

Gregory V, Bennett M, Orkhan M, Hajjar SA, Varsano N, Mendelson E, Zambon M, Ellis J, Hay A, Lin Y. 2002. Emergence of influenza A H1N2 reassortant viruses in the human population during 2001. *Virology* 300:1–7.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.

Han GZ, Boni MF, Li SS. 2010. No observed effect of homologous recombination on influenza C virus evolution. *Virol J.* 7:227.

Hasegawa M, Kishino H, Yano TA. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22:160–174.

Hay AJ, Gregory V, Douglas AR, Lin YP. 2001. The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci.* 356: 1861–1870.

Hedrick PW, Thomson G. 1986. A two-locus neutrality test: applications to humans, *E. coli* and lodgepole pine. *Genetics* 112:135–156.
Hoffmann E, Mahmood K, Yang CF, Webster RG, Greenberg HB, Kemble G. 2002. Rescue of influenza B virus from eight plasmids. *Proc Natl Acad Sci U S A.* 99:11411–11416.
Kanegae Y, Sugita S, Endo A, Ishida M, Senya S, Osako K, Nerome K, Oya A. 1990. Evolutionary pattern of the hemagglutinin gene of influenza B viruses isolated in Japan: cocirculating lineages in the same epidemic season. *J Virol.* 64:2860–2865.
Kawaoka Y, Krauss S, Webster RG. 1989. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *J Virol.* 63:4603–4608.
Lee HY, Chou JY, Cheong L, Chang NH, Yang SY, Leu JY. 2008. Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell* 135:1065–1073.
Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* 5:e1000520.
Lewontin RC. 1964. The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics* 49:49–67.
Lindstrom SE, Hiromoto Y, Nishimura H, Saito T, Nerome R, Nerome K. 1999. Comparative analysis of evolutionary mechanisms of the hemagglutinin and three internal protein genes of influenza B virus: multiple cocirculating lineages and frequent reassortment of the NP, m, and NS genes. *J Virol.* 73:4413–4426.
Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. 2011. Host phylogeny determines viral persistence and replication in novel hosts. *PLoS Pathog.* 7:e1002260.
Lu L, Lycett SJ, Brown AJL. 2014. Reassortment patterns of avian influenza virus internal segments among different subtypes. *BMC Evol Biol.* 14:16.
Lycett SJ, Baillie G, Coulter E, Bhatt S, Kellam P, McCauley JW, Wood JL, Brown IH, Pybus OG, Leigh Brown AJ Combating Swine Influenza Initiative-COSI Consortium. 2012. Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. *J Gen Virol.* 93: 2326–2336.
Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol.* 25:1459–1471.
Muller H. 1942. Isolating mechanisms, evolution and temeperature. *Biol Symp.* 6:71–125.
Nakagawa N, Nukuzuma S, Haratome S, Go S, Nakagawa T, Hayashi K. 2002. Emergence of an influenza B virus with antigenic change. *J Clin Microbiol.* 40:3068–3070.
Nerome R, Hiromoto Y, Sugita S, Tanabe N, Ishida M, Matsumoto M, Lindstrom SE, Takahashi T, Nerome K. 1998. Evolutionary characteristics of influenza B virus since its first isolation in 1940: dynamic circulation of deletion and insertion mechanism. *Arch Virol.* 143: 1569–1583.
O'Brien JD, Minin VN, Suchard MA. 2009. Learning to count: robust estimates for labeled distances between molecular sequences. *Mol Biol Evol.* 26:801–814.
Osterhaus ADME, Rimmelzwaan GF, Martina BEE, Bestebroer TM, Fouchier RAM. 2000. Influenza B virus in seals. *Science* 288: 1051–1053.
Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nat Rev Genet.* 11:175–180.
Rambaut A, Suchard M, Drummond A 2009. Tracer v1.5. [Internet]. Available from: http://tree.bio.ed.ac.uk/software/tracer/.
Reed C, Meltzer MI, Finelli L, Fiore A. 2012. Public health impact of including two lineages of influenza B in a quadrivalent seasonal influenza vaccine. *Vaccine* 30:1993–1998.
Rota PA, Wallis TR, Harmon MW, Rota JS, Kendal AP, Nerome K. 1990. Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983. *Virology* 175:59–68.
Ruane S, Pyron RA, Burbrink FT. 2011. Phylogenetic relationships of the cretaceous frog Beelzebufo from Madagascar and the placement of fossil constraints based on temporal and phylogenetic evidence. *J Evol Biol.* 24:274–285.
Shaw MW, Xu X, Li Y, Normand S, Ueki RT, Kunimoto GY, Hall H, Klimov A, Cox NJ, Subbarao K. 2002. Reappearance and global spread of variants of influenza B/Victoria/2/87 lineage viruses in the 2000–2001 and 2001–2002 seasons. *Virology* 303:1–8.
Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, et al. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459:1122–1125.
Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* 3:e285.
Ward MJ, Lycett SJ, Kalish ML, Rambaut A, Brown AJL. 2013. Estimating the rate of intersubtype recombination in early HIV-1 group M strains. *J Virol.* 87:1967–1973.
World Health Organization. 2009. Influenza fact sheet. Available from: http://www.who.int/mediacentre/factsheets/fs211/en/.
Zhao H, Nettleton D, Soller M, Dekkers JCM. 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet Res.* 86: 77–87.